Sparse sign-consistent Johnson–Lindenstrauss matrices: Compression with neuroscience-based constraints

Zeyuan Allen-Zhu, Rati Gelashvili, Silvio Micali[1], and Nir Shavit

Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139

Contributed by Silvio Micali, October 7, 2014 (sent for review February 13, 2014)

Johnson–Lindenstrauss (JL) matrices implemented by sparse random synaptic connections are thought to be a prime candidate for how convergent pathways in the brain compress information. However, to date, there is no complete mathematical support for such implementations given the constraints of real neural tissue. The fact that neurons are either excitatory or inhibitory implies that every so implementable JL matrix must be *sign consistent* (i.e., all entries in a single column must be either all nonnegative or all nonpositive), and the fact that any given neuron connects to a relatively small subset of other neurons implies that the JL matrix should be *sparse*. We construct sparse JL matrices that are sign consistent and prove that our construction is essentially optimal. Our work answers a mathematical question that was triggered by earlier work and is necessary to justify the existence of JL compression in the brain and emphasizes that inhibition is crucial if neurons are to perform efficient, correlation-preserving compression.

Johnson–Lindenstrauss compression | synaptic-connectivity matrices | sign-consistent matrices

1. Introduction

The existence of some form of compression in the brain is well accepted among neurobiologists. Its biological "evidence" proceeds from the brain's numerous convergent pathways, where information coming from a large number of neurons must be compressed into a small number of axons or neurons. Classical examples are the optic nerve fibers that carry information about the activity of 100 times as many photoreceptors* (1) or the pyramidal tract fibers that carry information from the (orders-of-magnitude) larger motor cortex to the spinal cord (2).

As far back as 1961, Barlow (3) hypothesized that the role of early sensory neurons is to remove statistical redundancy in sensory input. This "efficient encoding" theory has been studied by many, as surveyed in depth by Simoncelli and Olshausen (4).

A recent survey by Ganguli and Sompolinsky (5) highlights the importance of compression and compressed sensing in the neural system for reducing the dimensionality of the activity pattern. A fundamental question they pose is "How much can a neural system reduce the dimensionality of its activity patterns without incurring a large loss in its ability to perform relevant computations?" They identify, as a minimal requirement, the importance of preserving the similarity structure of the neuronal representations at the source area, to capture the idea (6, 7) that in higher perceptual or association areas in the brain, semantically similar objects elicit similar neural activity patterns.

Ganguli and Sompolinsky suggest that such compression can be achieved in the brain via random synaptic-connectivity matrices implementing *Johnson–Lindenstrauss* (*JL*) matrices. However, because each neuron is either excitatory or inhibitory, an additional constraint, *sign consistency*, is necessary for this implementation to work.

In this paper, we show for the first time to our knowledge that JL matrices can be simultaneously *compression efficient*, *sparse*, and *sign consistent* and are thus implementable by biologically plausible neural networks.

JL Compression and Synaptic Connectivity. Informally, JL compression (8) uses a random matrix A to map a long vector of reals, x, the *input*, to a much shorter vector of reals, $y = Ax$, the JL *output*. The JL result shows that if the number of input vectors one may ever need to compress is reasonably upper bounded, then the following property is satisfied:

Inner-product preservation. $\langle x, x' \rangle \approx \langle Ax, Ax' \rangle$ for all envisaged x and x'.†

Note that inner-product preservation implies the aforementioned "similarity property" of biological interest; that is,

Correlation preservation.

$$\frac{\langle x, x' \rangle}{\|x\| \cdot \|x'\|} \approx \frac{\langle Ax, Ax' \rangle}{\|Ax\| \cdot \|Ax'\|} \text{ for all envisaged } x \text{ and } x'.\ddagger$$

That is, similar JL inputs correspond to similar JL outputs.

The mentioned insight for implementing JL compression in the brain is random synaptic connectivity. An $m \times d$ JL matrix A is biologically implemented via the synaptic connections among (the *axons* of) d "input" neurons and (the *dendrites* of) $m < d$ "output" neurons. In essence,

Significance

Significant biological evidence indicates that the brain may perform some form of compression. To be meaningful, such compression should preserve pairwise correlation of the input data. It is mathematically well known that multiplying the input vectors by a sparse and fixed random matrix *A* achieves the desired compression. But, to implement such an approach in the brain via a synaptic-connectivity matrix, *A* should also to be sign consistent: that is, all entries in a single column must be either all nonnegative or all nonpositive. This is so because most neurons are either excitatory or inhibitory. We prove that sparse sign-consistent matrices can deliver the desired compression, lending credibility to the hypothesis that correlation-preserving compression occurs in the brain via synaptic-connectivity matrices.

Author contributions: S.M. and N.S. designed research; Z.A.-Z. and R.G. performed research; Z.A.-Z. and R.G. contributed new reagents/analytic tools; and S.M. and N.S. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: silvio@csail.mit.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1419100111/-/DCSupplemental.

*Tan C (2013) From vision to memory to concept learning. *Lecture*, MIT CBCL-CSAIL Brains, Minds and Machines Seminar Series.

†As usual, $\langle x, x' \rangle$ represents the inner product of x and x', that is, $\sum_i x_i x'_i$. Inner-product preservation immediately implies (and is in fact equivalent to) *norm preservation*: Namely, $\|x\| \approx \|Ax\|$ for all envisaged x.

‡As usual, $\|x\|$ represents the ℓ_2 norm of x, that is, $\sqrt{\langle x, x \rangle}$.

in a synaptic-connection matrix, the jth column corresponds to the connections of the jth input neuron, n_j. An entry (i,j) is 0 if n_j does not connect to the ith output neuron; else, it is the strength of their synaptic connection.

The encouraging aspect of the above biological implementation of a JL matrix A is that the random structure of A matches the randomness of neural connections.[§]

1.1. Three Interrelated Challenges.

Sign consistency. Let us explain the constraint in reference 16 of Rajan, Abbott, and Sompolinsky (10). According to Dale's principle, almost all neurons have one of the following two types: *excitatory* or *inhibitory*, but not both. The type of a neuron n essentially determines the "sign of the signal" it can transmit to a postsynaptic neuron p. As a standard approximation, an excitatory neuron n can only increase the activity of p, and an inhibitory one can only decrease it. Thus, a synaptic-connection matrix must be sign consistent (10, 11). That is, the nonzero entries of a column j must be either (*i*) *all positive*, if the jth input neuron n_j is excitatory, or (*ii*) *all negative*, if n_j is inhibitory.

Unfortunately, typical JL matrices are *not* sign consistent.

Sparsity. Let us emphasize another fundamental biological constraint: sparsity. A neuron may be connected to up to a few thousand postsynaptic neurons (12). (Furthermore, two neurons typically share multiple connections.) Thus, no synaptic-connectivity matrix could implement a dense JL matrix when m is large.[¶]

As originally constructed, JL matrices were dense. Sparse JL matrices have been recently constructed (14–18), but they are far from being sign consistent. Therefore, although the sign consistency of synaptic action may have a few exceptions, the extent to which the above mathematical constructions may be biologically relevant is not clear.

Efficiency. As we mentioned at the start of *Section 1, Introduction*, implementing an $m \times d$ JL matrix in the brain is interesting only if m is significantly smaller than d. [Of course, achieving such efficiency is more challenging with sign consistency, but Rajan and Abbott (10) have expressed optimism about the general ability to satisfy the latter constraint.]

Three prior approaches. Let us explain why these three challenges have not been simultaneously met.

A first and simplest way for JL matrices to be sign consistent is for them not to have any negative entries, corresponding to synaptic-connectivity matrices without inhibitory neurons. However, it is not hard to prove that nonnegative JL matrices must be *extremely* inefficient (e.g., $m \geq d/2$ for typical choices of parameters, *SI Appendix D*). This strong lower bound actually provides additional evidence for the cruciality of inhibition for neural functions.

The result of Rajan and Abbott (10) on the eigenvalue spectra of square matrices implies a way to transform JL matrices into sign-consistent ones (subject to mild assumptions on the inputs). However, the sign-consistent JL matrices they obtained were very dense: Half of their entries had to be nonzero (details in *Section 2, Related Mathematical Work*).

A third approach to sign-consistent JL matrices is implicitly provided by a transformation of Krahmer and Ward (19). Indeed, when applied to nonnegative matrices satisfying the restricted isometry property, their transformation yields sign-consistent JL matrices, but this construction can be proved to be much less efficient than ours.

[§]Although it may be easier to biologically construct a large random matrix, billions of years of evolution may not suffice for the emergence of a very special and very large matrix of neural connections. Moreover, this random construction need not be first found by evolution and then preserved genetically. That is, a good matrix A need not be the same across different individuals of the same species. It suffices that our DNA ensures that each individual, during development, randomly constructs *his own* matrix A.

[¶]Moreover, even if m were small—e.g., $m = 1{,}000$—it seems hard to find in the brain a complete bipartite graph with d "inputs" and m "outputs" (ref. 13, p. 35).

1.2. Our Contributions.

The mentioned biological constraints motivate the following purely mathematical question: *How efficient can sparse (randomly constructed) and sign-consistent JL matrices be?*

We answer this question exactly by providing tight upper and lower bounds.

We begin by formally stating the classical JL lemma (using norms rather than inner products):

Letting $m = \Theta(\varepsilon^{-2}\log(1/\delta))$, there exists a distribution $\mathcal{A}$ over $m \times d$ matrices such that, for any $x \in \mathbb{R}^d$, with probability at least $1-\delta$, $\|Ax\|_2 = (1\pm\varepsilon)\|x\|_2$.

The parameter ε measures the *distortion* introduced by the JL compression; in particular, one may consider $\varepsilon = 10\%$ (5). The parameter δ measures the *confidence* with which the $1\pm\varepsilon$ distortion is guaranteed. Because one may only need to compress polynomially many (rather than exponentially many) vectors in his lifetime, one typically chooses $\delta = 1/\mathbf{poly}(d)$ and thus $\log(1/\delta) = O(\log d)$. [With this choice of δ, after applying union bound, a matrix A generated from $\mathcal{A}$ is capable of compressing $\mathbf{poly}(d)$ envisaged vectors from $\mathbb{R}^d$, with high confidence.]

We prove two main results:

Theorem 1. "A Construction of Sparse, Efficient, and Sign-Consistent JL Matrices." *That is, letting $m = \Theta(\varepsilon^{-2}\log^2(1/\delta))$, there exists a distribution $\mathcal{A}$ over $m \times d$ sign-consistent matrices such that, for any $x \in \mathbb{R}^d$, with probability at least $1-\delta$, $\|Ax\|_2 = (1\pm\varepsilon)\|x\|_2$.*

More precisely, a matrix A generated according to $\mathcal{A}$ enjoys the following properties:

- *Sparsity*: Each column has $\Theta(\varepsilon^{-1}\log(1/\delta))$ nonzero entries.
- *Same magnitude*: All nonzero entries have the same absolute value.
- *Simplicity*: The positions of the nonzero entries in a column and the sign of a column itself are both randomly selected, independent from other columns.

Note that *Theorem 1* shows the norm preservation up to a multiplicative error $1\pm\varepsilon$. This implies correlation preservation up to an additive error $\pm O(\varepsilon)$; that is,

$$\frac{\langle x, x'\rangle}{\|x\|\cdot\|x'\|} = \frac{\langle Ax, Ax'\rangle}{\|Ax\|\cdot\|Ax'\|} \pm O(\varepsilon).$$

Note also that we cannot hope for a multiplicative error on correlation preservation because the correlation value is between −1 and 1, and thus a multiplicative error would imply the ability to recover orthogonal vectors (i.e., vectors with correlation zero) *precisely*.

Theorem 2. "Output-Length Optimality Among All Sign-Consistent JL Matrices." *That is, let $\mathcal{A}$ be a distribution over $m \times d$ sign-consistent matrices such that, for any $x \in \mathbb{R}^d$, with probability at least $1-\delta$, $\|Ax\|_2 = (1\pm\varepsilon)\|x\|_2$. Then, $m = \tilde{\Omega}(\varepsilon^{-2}\log(1/\delta)\cdot\min\{\log d, \log(1/\delta)\})$.*[#]

Note that in the interesting parameter regime of $\delta = 1/\mathbf{poly}(d)$, our lower bound becomes $\tilde{\Omega}(\varepsilon^{-2}\log^2(1/\delta))$; that is, it essentially matches our upper bound.

1.3. In Sum.

Our work closes an open mathematical question that is necessary to justify the existence of JL compression in the brain. Our work provides the missing support by constructing JL matrices that are simultaneously sparse and sign consistent and offer the most efficient JL compression possible; moreover, our work interestingly implies that inhibition is crucial if neurons are to perform efficient, correlation-preserving compression.

[#]Recall that the notation of $\tilde{\Omega}(N)$ signifies that logarithmic factors of N are ignored. Thus, in our case, factors of $\log(1/\varepsilon)$ and $\log\log(1/\delta)$ are ignored in this lower bound.

Looking forward, the brain has inspired several models of computation, from perceptrons (20) to neural networks (21), which have already proved fruitful in many fields and in machine learning in particular.

Computer scientists have started studying computational models that are increasingly biologically relevant, for fundamental tasks such as concept representation and formation (22) and memory allocation (23–25). We consider our paper a further step in this direction.

2. Related Mathematical Work

To the best of our knowledge, the only mathematical analysis of a random sign-consistency matrix is the one suggested by Rajan and Abbott (10) and followed by ref. 11. Although their results are about the eigenvalue spectra of a random sign-consistent *square* matrix, it implies[‖] the following way of constructing an $m \times d$ sign-consistent JL matrix A'.

- First, construct an $m \times d$ JL matrix A, by randomly assigning each entry of A from $\{-1/\sqrt{m}, 1/\sqrt{m}\}$.
- Second, construct an $m \times d$ special matrix M, by assigning each entry of (a random) half of the columns of M to be $-1/\sqrt{m}$ and each entry of the remaining half to be $1/\sqrt{m}$.
- At last, set $A' \stackrel{\text{def}}{=} (1/\sqrt{2})(A+M)$.

Then, A' is sign consistent, and $A'x = Ax$ (so A' is JL), assuming that x satisfies $\sum_{i \in [d]} x_i = 0$. This assumption aside, however, the resulting matrix A' must be very dense.

The classical JL construction requires a distribution over *dense* matrices (e.g., i.i.d. Gaussian or Rademacher entries), but achieves a target dimension of $m = O(\varepsilon^{-2} \log(1/\delta))$, which is essentially optimal (26). A beautiful line of work (19, 27–31) has made use of the Hadamard or Fourier matrices in the JL construction, to speed up the matrix–vector multiplication to nearly linear time. However, their matrices are *dense too*. Recent constructions (14–18) yield sparse JL matrices that have $O(\varepsilon^{-1} \log(1/\delta))$ rows, which have been shown to be essentially optimal (32).

Although not applicable to JL matrices, Clarkson and Woodruff have shown how to construct sign-consistent and *optimally sparse* (namely, a single nonzero entry per column) random matrices (33). Their matrices preserve correlation for inputs satisfying an algebraic constraint, namely, coming from a *hidden subspace*. By contrast, we want to compress arbitrary inputs.

For numerous applications of JL compression in computer science, see refs. 34 and 35.

A JL matrix A can be easily constructed (with very high probability) by choosing each entry at random. Of course, given such a randomly constructed matrix, it would be nice to reconstruct, with meaningful approximation, the original JL-compressed input x from Ax;[**] but this cannot be done without assuming that the inputs are of a *restricted type* [e.g., close to vectors with few nonzero entries (36, 37)]. However, even without reconstructing the inputs, inner-product preservation allows us to perform a variety of fundamental computations on the JL outputs directly and thus with great efficiency. This includes nearest neighbors (38), classification (39), regression (40), and many others.

3. A Simple Experimental Illustration

Let us consider a simple experiment to numerically verify the dependency $m = \Theta(\varepsilon^{-2} \log(1/\delta))$ in the classical JL construction and the dependency $m = \Theta(\varepsilon^{-2} \log^2(1/\delta))$ in our novel construction. Rather than fixing the distortion ε and the confidence δ and computing the target dimension m, we find it more convenient to fix m and δ first and then compute ε.

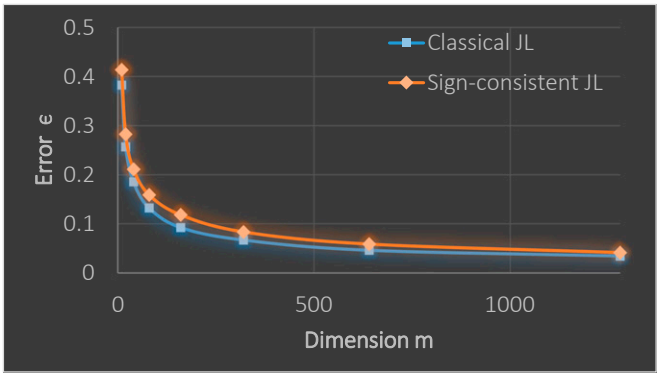

Fig. 1. The classical and our sign-consistent JL constructions.

Specifically, we fix $d = 3{,}000$ and $\delta = 0.1$. Consider the following values of m: $m = 10, 20, 40, \ldots, 1{,}280$; and then numerically compute the distortion ε for each value of m, both for the classical and for our new JL construction.

In the classical JL construction, the $m \times d$ dimension matrix A is chosen so that each entry is either $1/\sqrt{m}$ or $-1/\sqrt{m}$, each with half probability.

In our construction, we first define the column sparsity $s = \lfloor \sqrt{m} \rfloor$. [Note that this is consistent with the parameters suggested by *Theorem 1*: $s = \Theta(\varepsilon^{-1} \log(1/\delta)) = \Theta(\sqrt{m})$.] Then, for each column of A, we randomly choose s entries of this column and then flip a fair coin: If it is heads, we set each of these s entries to $1/\sqrt{s}$; if it is tails, we set each of them to $-1/\sqrt{s}$.

For the above two constructions, we apply the JL transformation Ax on 1,000 randomly chosen inputs $x_1, \ldots, x_{1{,}000} \in \mathbb{R}^d$. For each construction, we compute the 1,000 distortions $\left| \|Ax_i\|_2 / \|x_i\|_2 - 1 \right|$, call ε the 100th highest distortion, and plot ε in Fig. 1. (This process is equivalent to choosing $\delta = 0.1$ and throwing out the highest δ fraction of the distortions. Indeed, $100 = \delta \cdot 1{,}000$.)

The experiment illustrates that for both curves, whenever m is enlarged by a factor of 4, the error ε decreases approximately by a factor of 2. This corresponds to the dependency $m \propto \varepsilon^{-2}$ in both constructions. Also note that the blue curve falls slightly below the red curve, corresponding to the difference between the dimension choice of $m = \Theta(\varepsilon^{-2} \log(1/\delta))$ in the classical JL construction and $m = \Theta(\varepsilon^{-2} \log(1/\delta)^2)$ in ours.

4. Proof Sketch of Theorem 1

Let $\mathcal{A}^{m,d,s}$ be the distribution of $m \times d$ matrices defined as follows. For each of the d columns, we choose uniformly at random s distinct entries [of $\binom{m}{s}$ possibilities] and assign a random value between $\{-1/\sqrt{s}, 1/\sqrt{s}\}$ (with half probability each) to these s entries, while leaving it zero in other entries of the same column.[††]

Theorem 1. *Letting $m = \Theta(\varepsilon^{-2} \log^2(1/\delta))$ and $s = \Theta(\varepsilon^{-1} \log(1/\delta))$, for any $x \in \mathbb{R}^d$, with probability at least $1-\delta$, we have $\|Ax\|_2 = (1 \pm \varepsilon)\|x\|_2$ over the choice of $A \sim \mathcal{A}^{m,d,s}$.*

The proof of *Theorem 1* is quite complex and is given in *SI Appendix A*. (In particular, the classical technique of the Hanson–Wright inequality fails to give a tight upper bound in our case, just like ref. 18.) Below, we outline just the important ingredients of the proof.

[‖] In fact, given any random matrix whose eigenvalues are randomly distributed on the complex unit disk, a random subset of its rows forms a JL matrix.

[**] To be sure, inner-product preservation always implies a weak form of reconstructability. Namely, each entry x_i of an input vector x can be reconstructed up to an additive error of $\varepsilon \cdot \|x\|_2$.

[††] Our theorem remains true if one divides each column into $\lceil m/s \rceil$ blocks and chooses one random entry from each block and/or if one uses $\Theta(\log(1/\delta))$-wise independent hash functions to generate $\mathcal{A}^{m,d,s}$.

Proof Sketch. Observe that the entries of a matrix $A \in \mathbb{R}^{m\times d}$ that we construct can be written as $A_{i,j} = \eta_{i,j}\sigma_j/\sqrt{s}$, where $\sigma_j \in \{-1,1\}$ is chosen uniformly at random, and $\eta_{i,j} \in \{0,1\}$ is an indicator variable for the event $A_{i,j} \neq 0$. All of the $\{\sigma_j\}_{j\in[d]}$ are independent; $\{\eta_{i,j}\}_{i\in[m],j\in[d]}$ are independent across columns, but not independent (and in fact negatively correlated) in the same column, because there are exactly s nonzero entries per column.

Given any fixed $x \in \mathbb{R}^d$ with $\|x\|_2 = 1$, let us study the following random variable:

$$Z \stackrel{\text{def}}{=} \|Ax\|_2^2 - 1 = \frac{1}{s}\cdot \sum_{r=1}^{m}\sum_{i\neq j\in[d]} \eta_{r,i}\eta_{r,j}\sigma_i\sigma_j x_i x_j.$$

To show that $|Z| \le \varepsilon$ with probability at least $1-\delta$, we need a good upper bound on the tth moment of Z [note that we will eventually choose $t = \Theta(\log(1/\delta))$]:

$$s^t\cdot\mathbb{E}[Z^t] = \sum_{\substack{i_1,\dots,i_t,j_1,\dots,j_t\in[d]\\ i_1\neq j_1,\dots,i_t\neq j_t}} \left(\prod_{u=1}^{t} x_{i_u}x_{j_u}\right)\left(\mathop{\mathbb{E}}_{\sigma}\prod_{u=1}^{t}\sigma_{i_u}\sigma_{j_u}\right) \times\left(\mathop{\mathbb{E}}_{\eta}\prod_{u=1}^{t}\sum_{r=1}^{m}\eta_{r,i_u}\eta_{r,j_u}\right).$$

The authors of ref. 18 analyzed a similar expression but with $\sigma_{i_u}\sigma_{j_u}$ replaced by $\sigma_{r,i_u}\sigma_{r,j_u}$. In their case, they decompose Z into subexpressions $Z = Z_1 + \cdots + Z_m$: Each Z_r contains all terms with the same row r (e.g., $\sigma_{r,\star}$ and $\eta_{r,\star}$) and can be analyzed separately. This greatly simplifies their job because Z_r and $Z_{r'}$ are negatively correlated when $r \neq r'$. In contrast, if we did the same thing, we would not have the same negative correlation anymore: Z_r and $Z_{r'}$ both contain the same random variables σ_i for all $i \in [d]$. Thus, we have to analyze the whole expression at once.

Reusing ideas from refs. 16–18, we analyze Z^t by associating monomials that appear in Z^t to *directed multigraphs with labeled edges*: An $x_{i_u}x_{j_u}$ term corresponds to a directed edge with label u from vertex i_u to vertex j_u. We then group the monomials together based on their associated graphs and prove the following lemma. (Its proof is analogous to that of equation 13 in ref. 19 but is more tedious.)

Lemma 1. $s^t\cdot\mathbb{E}[Z^t] \le e^t\sum_{v=2}^{t}\sum_{G\in\mathcal{G}''_{v,t}}\left((1/t^t)\prod_{p=1}^{v}\sqrt{d_p}^{\,d_p}\right)\cdot\sum_{r_1,\dots,r_t\in[m]}\prod_{i=1}^{w}(s/m)^{v_i}$. *Here,*

- $\mathcal{G}''_{v,t}$ *is a set of directed multigraphs with* v *labeled vertices* (1 *to* v) *and* t *labeled edges* (1 *to* t).
- d_p *is the total degree of vertex* $p \in [v]$ *in a graph* $G \in \mathcal{G}''_{v,t}$.[‡‡]
- w *and* $v_1,\dots,v_w$ *are defined by* G *and* $r_1,\dots,r_t$ *as follows. Let an edge* $u \in [t]$ *be colored with* $r_u \in [m]$; *then we define* w *to be the number of distinct colors used in* $r_1,\dots,r_t$ *and* v_i *to be the number of vertices incident to an edge with color* $i \in [w]$.

As one may have observed, for the aforementioned reason, we need to deal with many rows (e.g., row $r_1,\dots,r_t$) together, introducing a concept of *color* defined above. To be precise, a directed edge (i_u, j_u) is now also colored with $r_u \in [m]$, and this is a major difference between our *Lemma 1* and for instance equation 13 in ref. 19. In essence, we are dealing with three-dimensional tuples (i_u, j_u, r_u) rather than just (i_u, j_u).

This difference is critical for obtaining a tight bound for Z^t: We have to bound the $\prod_{p=1}^{v}\sqrt{d_p}^{\,d_p}$ terms separately for graphs of different colors (as otherwise we will lose a $\log(1/\delta)$ factor in the proof). In other words, instead of enumerating $G \in \mathcal{G}''_{v,t}$ as a whole, we now have to enumerate subgraphs of different colors separately and then combine the results. Below is one way (and perhaps the only way we believe) to enumerate G that can lead to tight upper bounds:

$$s^t\cdot\mathbb{E}[Z^t] \le e^t \underbrace{\sum_{v=2}^{t}}_{i}\underbrace{\sum_{w=1}^{t}\binom{m}{w}}_{ii}\underbrace{\sum_{\substack{c_1,\dots,c_w\\ c_1+\dots+c_w=t\\ c_i\ge 1}}\binom{t}{c_1,\dots,c_w}}_{iii} \times\underbrace{\sum_{\substack{v_1,\dots,v_w\\ 2\le v_i\le 2c_i}}}_{iv}\left(\frac{s}{m}\right)^{v_1+\dots+v_w}\underbrace{\sum_{f_1,\dots,f_w}}_{v}\underbrace{\sum_{\forall i, G_i\in\mathcal{G}''_{v_i,c_i}}}_{vi}\frac{1}{t^t}\prod_{p=1}^{v}\sqrt{d_p}^{\,d_p}. \quad [4.1]$$

This gigantic expression enumerates all $G \in \mathcal{G}''_{v,t}$ and their colorings $r_1,\dots,r_t \in [m]$ in six steps:

i) Number of graph vertices, $v \in \{2,\dots,t\}$; the vertices are labeled by $1,2,\dots,v$.
ii) Number of used edge colors, $w \in \{1,\dots,t\}$, and all $\binom{m}{w}$ possibilities of choosing w colors.
iii) Edge colorings of the graph using selected w colors: How many (denoted by $c_i \ge 1$) edges are colored in color i and which of the t edges are colored in color i.
iv) Number of vertices $v_i \in \{2,\dots,2c_i\}$ in each G_i, the subgraph containing edges of color i.
v) All possible increasing functions $f_i : [v_i] \to [v]$, such that $f_i(j)$ maps vertex j in G_i to the $f_i(j)$ th global vertex. [And we ensure $f_i(j) < f_i(k)$ for $j < k$ to reduce double counting.]
vi) All graphs $G_i \in \mathcal{G}''_{v_i,c_i}$ with v_i labeled vertices (1 to v_i) and c_i labeled edges (1 to c_i). (Using all of the information above, d_p, the degree of vertex $p \in [v]$, is well defined.)

We emphasize here that any pair of graph $G \in \mathcal{G}''_{v,t}$ and coloring $r_1,\dots,r_t \in [m]$ will be generated *at least once* in the above procedure.[§§] Thus, [**4.1**] follows from *Lemma 1*, because the summation terms also have the same value $(s/m)^{v_1+\dots+v_w}(1/t^t)\prod_{p=1}^{v}\sqrt{d_p}^{\,d_p}$.

It is now possible to consider G_i s separately in [**4.1**] and prove the following lemma:

Lemma 2.

$$s^t\cdot\mathbb{E}[Z^t] \le 2^{O(t)}\sum_{v=2}^{t}\sum_{w=1}^{t}\binom{m}{w}\sum_{\substack{c_1,\dots,c_w\\ c_1+\dots+c_w=t\\ c_i\ge 1}}\binom{t}{c_1,\dots,c_w} \times\sum_{\substack{v_1,\dots,v_w\\ 2\le v_i\le 2c_i}}\prod_{j=1}^{w}\left(\frac{s}{m}\right)^{v_j}v_j^{c_j}\binom{v-1}{v_j-1}.$$

Some delicate issues arise here. For instance, one may use the Cauchy–Shwartz technique of ref. 18 to deduce

[‡‡]The total degree of a vertex is defined as the number of incident edges regardless of direction.

[§§]This follows from the fact that G and $r_1,\dots,r_t$ together determine (*i*) w, the number of used colors; (*ii*) G_i for each $i \in [w]$ (with v_i vertices and c_i edges), the subgraph of G of the ith used color; and (*iii*) f_i, the vertex mapping from G_i back to G. Any such triple will be generated at least once in [**4.1**]. Note also that we may have double counts but it will not affect our asymptotic upper bound.

$$\sum_{f_1,\dots,f_w}\sum_{\forall i,G_i\in\mathcal{G}''_{v_i,c_i}}\frac{1}{t^t}\prod_{p=1}^{v}\sqrt{d_p}^{\,d_p}\le\sum_{\substack{v_1,\dots,v_w\\2\le v_i\le 2c_i}}\prod_{j=1}^{w}v_j^{c_j}\binom{v}{v_j},$$

getting a weaker upper bound as it replaces $\binom{v-1}{v_j-1}$ with $\binom{v}{v_j}$ in *Lemma 2*. However, even such a simple replacement leads to a $\log(1/\delta)$ factor loss! Finally, after enduring layers of algebraic simplifications we prove the following:

Lemma 3. $s^t\cdot\mathbb{E}[Z^t]\le 2^{O(t)}\cdot t^t\left(\frac{s^2}{m}\right)^t$.

By *Lemma 3*, there exists a constant C such that $\mathbb{E}[Z^t]\le(Cts/m)^t$. Using Markov's inequality, we have $\Pr[|Z|>\varepsilon]\le\mathbb{E}[Z^t]/\varepsilon^t\le(Cts/\varepsilon m)^t$. We now set parameters $t\stackrel{\text{def}}{=}\log(1/\delta)$, $s\stackrel{\text{def}}{=}\varepsilon^{-1}t$, and $m\stackrel{\text{def}}{=}\varepsilon^{-1}ts/2C$. Plugging them in we get $\Pr[|Z|>\varepsilon]\le\delta$ as desired, finishing the proof of *Theorem 1*. ■

5. Statement of Theorem 2

Here we formally state our second theorem and defer its full proof to *SI Appendix B*.

Theorem 2. *There is some fixed $\varepsilon_0\in(0,1/2)$ such that for all $\varepsilon\in(1/\sqrt{d},\varepsilon_0)$, all $m\le O(d/\log(1/\varepsilon))$, and all $\delta\le\varepsilon^{12}$, the following holds. Let $\mathcal{A}$ be a distribution over $m\times d$ sign-consistent matrices such that, for any $x\in\mathbb{R}^d$, with probability at least $1-\delta$, the ℓ_2 embedding $\|Ax\|_2=(1\pm\varepsilon)\|x\|_2$ has ε distortion. Then,*

$$m=\Omega\left(\frac{\varepsilon^{-2}\log(1/\delta)}{\log(\varepsilon^{-2}\log(1/\delta))}\min\left\{\log d,\log\frac{1}{\delta}\right\}\right).$$

1. Sterling P, Demb JB (2004) Retina. *The Synaptic Organization of the Brain*, ed Shepherd GM (Oxford Univ Press, Oxford), 5th Ed, Chap 6.
2. Ghez C, Wolpert DM, Pearson K (2012) The organization and planning of movement. *Principles of Neural Science*, eds Kandel ER, Schwartz JH, Jessell TM (McGraw-Hill, New York), 5th Ed, Chap 33.
3. Barlow HB (1961) Possible principles underlying the transformation of sensory messages. *Sensory Communication*, ed Rosenblith WA (MIT Press, Cambridge, MA), Chap 13.
4. Simoncelli EP, Olshausen BA (2001) Natural image statistics and neural representation. *Annu Rev Neurosci* 24(1):1193–1216.
5. Ganguli S, Sompolinsky H (2012) Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis. *Annu Rev Neurosci* 35:485–508.
6. Kiani R, Esteky H, Mirpour K, Tanaka K (2007) Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *J Neurophysiol* 97(6):4296–4309.
7. Rogers TT, McClelland JL (2004) *Semantic Cognition: A Parallel Distributed Processing Approach* (MIT Press, Cambridge, MA).
8. Johnson WB, Lindenstrauss J (1984) Extensions of Lipschitz mappings into a Hilbert space. *Contemp Math* 26:189–206.
9. Rajan K, Abbott LF, Sompolinsky H (2010) Stimulus-dependent suppression of chaos in recurrent neural networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 82(1 Pt 1):011903.
10. Rajan K, Abbott LF (2006) Eigenvalue spectra of random matrices for neural networks. *Phys Rev Lett* 97(18):188104.
11. Gray RT, Robinson PA (2009) Stability and structural constraints of random brain networks with excitatory and inhibitory neural populations. *J Comput Neurosci* 27(1): 81–101.
12. Kandel ER, Schwartz JH, Jessell TM (2000) *Principles of Neural Science* (McGraw-Hill, New York).
13. Buzsaki G (2006) *Rhythms of the Brain* (Oxford Univ Press, Oxford).
14. Achlioptas D (2003) Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J Comput Syst Sci* 66(4):671–687.
15. Dasgupta A, Kumar R, Sarlós T (2010) A sparse Johnson-Lindenstrauss transform. *Proceedings of the 42nd ACM Symposium on Theory of Computing - STOC '10* (ACM, New York), pp 341–350.
16. Kane DM, Nelson J (2010) A derandomized sparse Johnson-Lindenstrauss transform. *Technical Report*. arXiv:1006.3585.
17. Braverman V, Ostrovsky R, Rabani Y (2010) Rademacher chaos, random Eulerian graphs and the sparse Johnson-Lindenstrauss transform. *Technical Report*. arXiv: 1011.2590.
18. Kane DM, Nelson J (2012) Sparser Johnson-Lindenstrauss transforms. *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms - SODA '12* (ACM and SIAM, New York), pp 1195–1206.
19. Krahmer F, Ward R (2011) New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property. *SIAM J Math Anal* 43(3):1269–1281.
20. Rosenblatt F (1958) The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol Rev* 65(6):386–408.
21. Fukushima K (1980) Neocognitron: A self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol Cybern* 36(4): 193–202.
22. Valiant LG (1994) *Circuits of the Mind* (Oxford Univ Press, New York).
23. Valiant LG (2005) Memorization and association on a realistic neural model. *Neural Comput* 17(3):527–555.
24. Feldman V, Valiant LG (2009) Experience-induced neural circuits that achieve high capacity. *Neural Comput* 21(10):2715–2754.
25. Valiant LG (2012) The hippocampus as a stable memory allocator for cortex. *Neural Comput* 24(11):2873–2899.
26. Alon N (2009) Perturbed identity matrices have high rank: Proof and applications. *Combin Probab Comput* 18(1–2):3–15.
27. Ailon N, Liberty E (2008) Fast dimension reduction using Rademacher series on dual BCH codes. *Discrete Comput Geom* 42(4):615–630.
28. Ailon N, Chazelle B (2009) The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM J Comput* 39(1):302–322.
29. Hinrichs A, Vybíral J (2011) Johnson-Lindenstrauss lemma for circulant matrices. *Random Structures & Algorithms* 39(3):391–398.
30. Vybíral J (2011) A variant of the Johnson-Lindenstrauss lemma for circulant matrices. *J Funct Anal* 260(4):1096–1105.
31. Ailon N, Liberty E (2013) An almost optimal unrestricted fast Johnson-Lindenstrauss transform. *ACM Trans Algorithms* 9(3):1–12.
32. Nelson J, Nguyễn HL (2013) Sparsity lower bounds for dimensionality reducing maps. *Proceedings of the 45th Annual ACM Symposium on Theory of Computing - STOC '13* (ACM, New York), pp 101–110.
33. Clarkson KL, Woodruff DP (2013) Low rank approximation and regression in input sparsity time. *Proceedings of the 45th Annual ACM Symposium on Theory of Computing - STOC '13* (ACM, New York), pp 81–90.
34. Indyk P (2001) Algorithmic applications of low-distortion geometric embeddings. *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science* (IEEE, Piscataway, NJ), pp 10–33.
35. Vempala SS (2004) *The Random Projection Method* (American Mathematical Society, Providence, RI), Vol 65.
36. Candes EJ, Tao T (2005) Decoding by linear programming. *IEEE Trans Inf Theory* 51(12):4203–4215.
37. Candes EJ, Romberg JK, Tao T (2006) Stable signal recovery from incomplete and inaccurate measurements. *Commun Pure Appl Math* 59(8):1207–1223.
38. Indyk P, Motwani R (1998) Approximate nearest neighbors: Towards removing the curse of dimensionality. *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing* (ACM, New York), pp 604–613.
39. Blum A (2005) Random projection, margins, kernels, and feature-selection. *Proceedings of the 2005 International Conference on Subspace, Latent Structure and Feature Selection* (Springer, Berlin), pp 52–68.
40. Zhou S, Lafferty J, Wasserman L (2009) Compressed and privacy-sensitive sparse regression. *IEEE Trans Inf Theory* 55(2):846–866.